Knowledge Distillation

Assignment 12

Submitted by

Ariful Islam

ID: 2010576130

Department of Computer Science

University of Rajshahi

Date: July 31, 2025

Course:

CSE4261

Instructor:

Dr. Sangeeta Biswas

1 Introduction

In this report, we will show and compare the performance of four classifiers on the CIFAR-10 dataset, which comprises 60,000 32×32 color images across 10 object categories. We begin by designing and training a compact CNN model from scratch to serve as our baseline. Next, we adapt two state-of-the-art CNN architectures—pre-trained on the ImageNet dataset—by fine-tuning their final layers for CIFAR-10 classification.

2 Question

Write a report by comparing performances of four 10 class classifiers after:

1. building a small CNN based classifier, and training it by your favorite dataset (except the MNIST digit or fashion dataset) having images of 10 classes.

2. preparing your two favorite big CNNs, pre-trained by ImageNet dataset for 1000 classes, for classifying 10 classes of your chosen dataset by fine tuning last 1/2 layers

3. transferring knowledge from your big fine-tuned one classifier to your own small CNN

4. transferring knowledge from your big fine-tuned two classifiers to your own small CNN

3 Methodology

3.1 Dataset and Preprocessing

Here, we use the CIFAR-10 dataset to perform all experiments throughout this study. CIFAR-10 consists of 60,000 color images categorized into 10 distinct classes, each with a resolution of 32×32 pixels. All datasets are normalized using the standard ImageNet mean and standard deviation. The validation and test sets use only normalization to maintain evaluation consistency.

Table 1: Dataset Summary

Split	**Size**
Training	9,000
Validation	1,000
Test	10,000
Image Size	32×32×3
Classes	10

3.2 Custom CNN Architecture

The custom CNN consists of two convolutional layers with 3×3 kernels, each followed by max-pooling, and two fully connected layers. The model processes 32×32 RGB images, reducing spatial dimensions from 32×32 to 8×8 with 32 channels before flattening. It has roughly 269,000 trainable parameters and a memory footprint of about 1.27 MB during training. This lightweight architecture balances efficiency and performance for CIFAR-10 classification.

Table 2: Model Summary

Layer (type)	**Output Shape**	**Param #**
Conv2d-1	[-1, 16, 32, 32]	448
MaxPool2d-2	[-1, 16, 16, 16]	0
Conv2d-3	[-1, 32, 16, 16]	4,640
MaxPool2d-4	[-1, 32, 8, 8]	0
Linear-5	[-1, 128]	262,272
Linear-6	[-1, 10]	1,290
Total params:		**268,650**
Trainable params:		**268,650**
Non-trainable params:		0

Memory Usage	**Size (MB)**
Input size	0.01
Forward/backward pass size	0.24
Params size	1.02
Estimated Total Size	1.27

3.3 Pre-Trained CNNs and Fine-Tuning

Two pre-trained models, ResNet50 and ResNet18, were used, originally trained on the ImageNet dataset for 1000 classes. For fine-tuning:

- **ResNet50 (Last Layer)**: The final fully connected layer was replaced with a new layer for 10 classes. Only this layer was trained, while earlier layers were frozen.
- **ResNet18 (Last Two Layers)**: The final fully connected layer and the last convolutional block were fine-tuned, with other layers frozen.

Fine-tuning was performed using the Adam optimizer with a learning rate of 0.0001, categorical cross-entropy loss, and a batch size of 32 for 20 epochs.

4 Results Discussion

4.1 For Task 1

The small CNN model trained on CIFAR-10 achieved steady improvements over 30 epochs, with validation accuracy increasing from 39.4 percent to 63.4 percent. The final test accuracy of

62.43 percent demonstrates the model's reasonable capability to classify the 10 classes despite its compact architecture and limited training data.

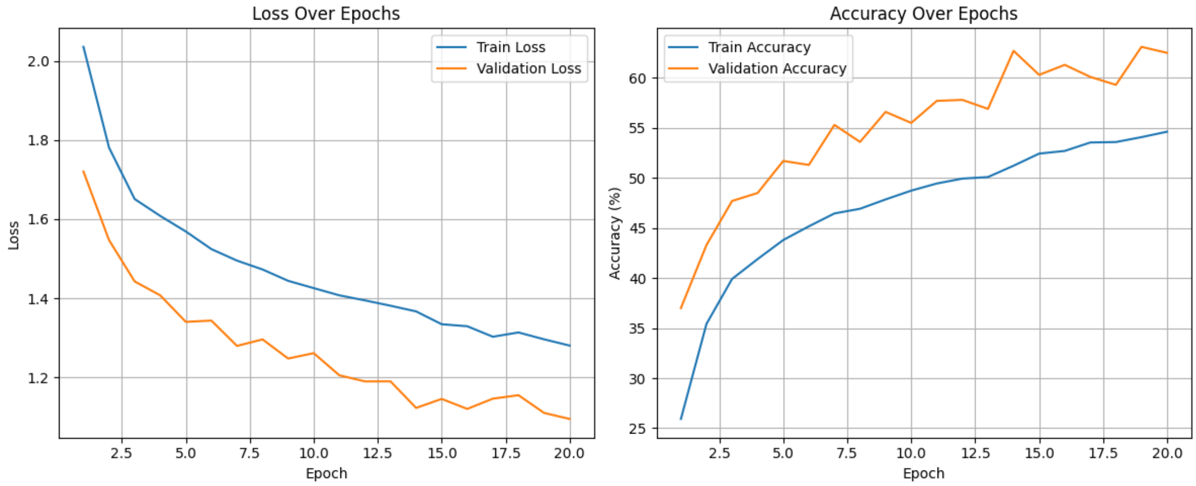


Figure 1: Accuracy and loss in training and validation

Table 3: Training and Evaluation Metrics

Metric	**Value**
Final Train Loss	1.1739
Final Train Accuracy	57.84%
Final Validation Loss	1.0587
Final Validation Accuracy	63.40%
Final Test Accuracy	62.43%

4.2 Task 2

ResNet50 and ResNet18, pre-trained on ImageNet, were fine-tuned on CIFAR-10 by adjusting their first convolution layers and retraining the deeper layers (layer3, layer4, and final classifier). ResNet50 achieved a validation accuracy of 80.2%, while ResNet18 reached 78.0%, both significantly outperforming the small custom CNN. This confirms that selective fine-tuning of large pre-trained models effectively leverages learned features for improved performance on smaller datasets.

Table 4: Comparison of our CNN and pretrained model

Model	**Train Loss**	**Train Accuracy**	**Val Loss**	**Val Accuracy**	**Test Accuracy**
Small CNN	1.1739	57.84%	1.0587	63.40%	62.43%
ResNet18	0.5757	79.99%	0.6967	78.00%	—
ResNet50	0.5292	81.18%	0.6040	80.20%	—

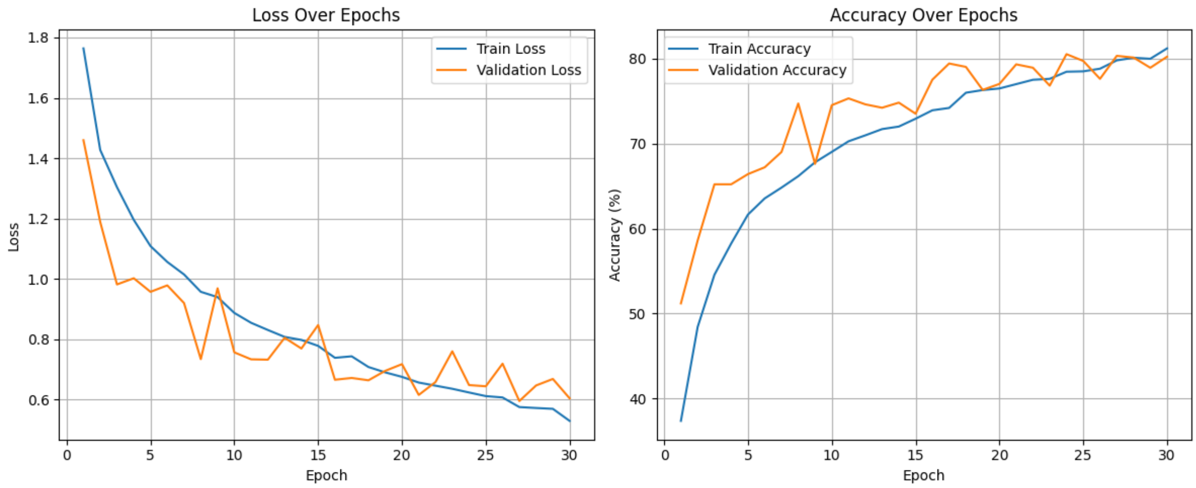


Figure 2: Accuracy and loss in training and validation in ResNet50

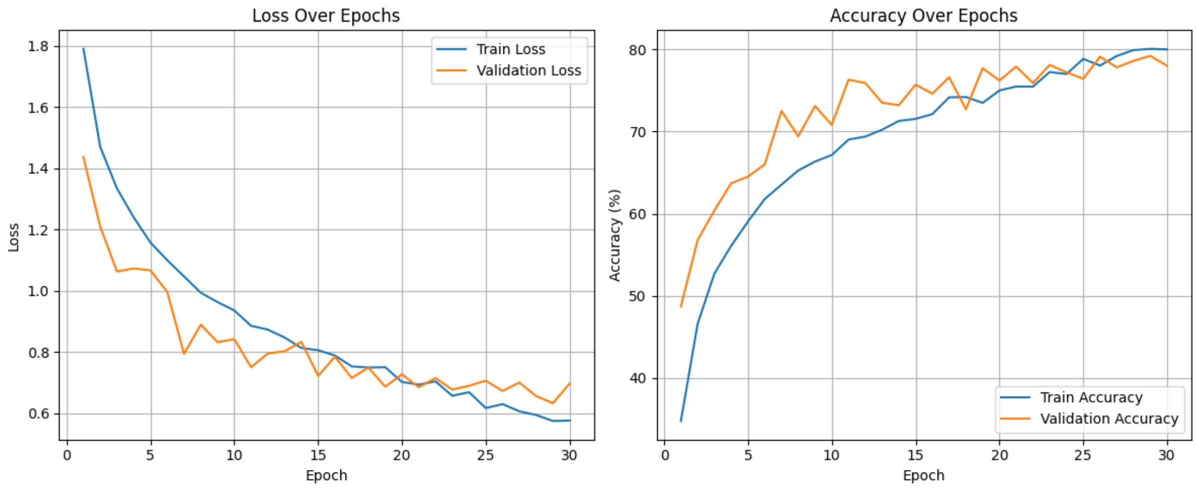


Figure 3: Accuracy and loss in training and validation in ResNet18

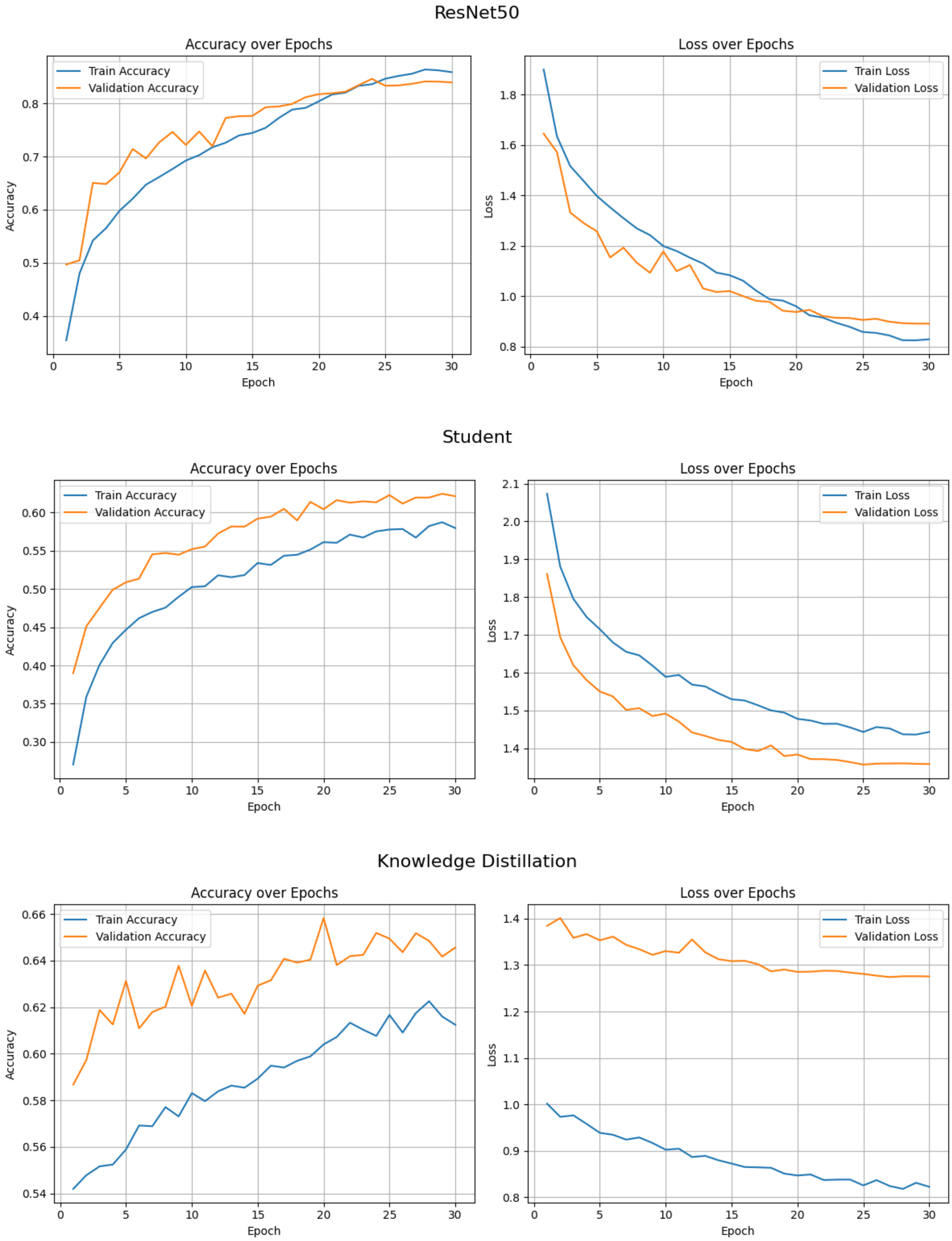
Task 3: Knowledge Transfer Approaches

In Task 3, knowledge is transferred from a large fine-tuned ResNet50 classifier to a smaller CNN model. Three training approaches were evaluated:

1. Fine-tuning ResNet50 (teacher model) on the target dataset to achieve high accuracy.
2. Training the small CNN directly on the target dataset.
3. Knowledge distillation, where the small CNN (student) learns from the soft outputs of the large ResNet50 (teacher) to improve its performance.

Observations

- The ResNet50 teacher model achieved the highest validation accuracy (approximately 83.98%), demonstrating strong performance due to its capacity and pre-training.
- The small CNN trained alone achieved moderate accuracy (approximately 62.25%), lim-



ited by its smaller capacity.

- The knowledge distillation approach improved the small CNN's performance significantly (approximately 66.34%), surpassing training from scratch by leveraging the teacher's knowledge.

Table 5: Performance Comparison of Different Models and Training Approaches

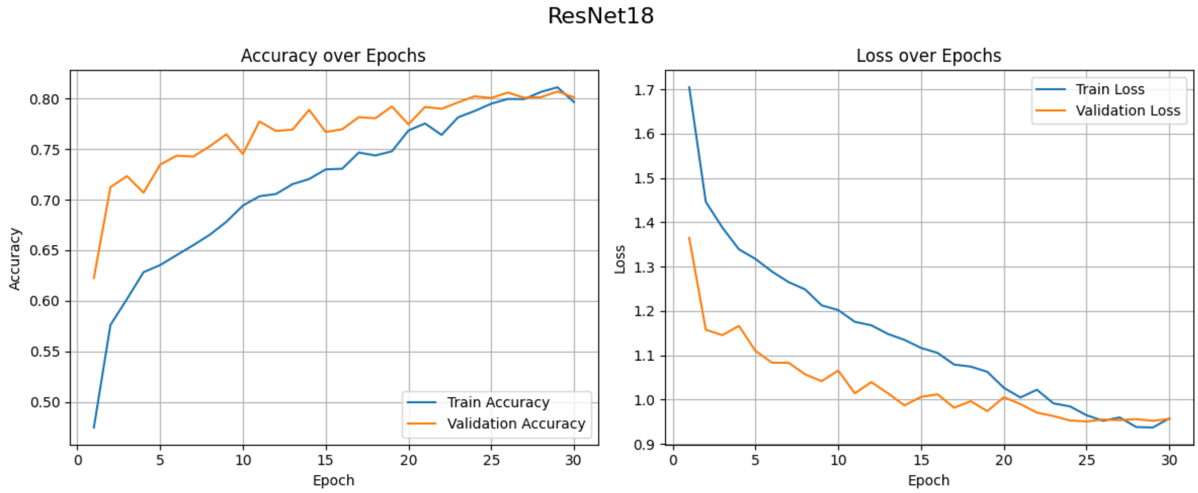
Model / Approach	Best Val Accuracy	Best Val Loss	Notes
ResNet50 (Teacher)	83.98%	0.8913	Large, fine-tuned model
Small CNN (From Scratch)	62.25%	1.3488	Small model trained directly
Small CNN (Distilled)	66.34%	1.2700	Small model trained with knowledge distillation

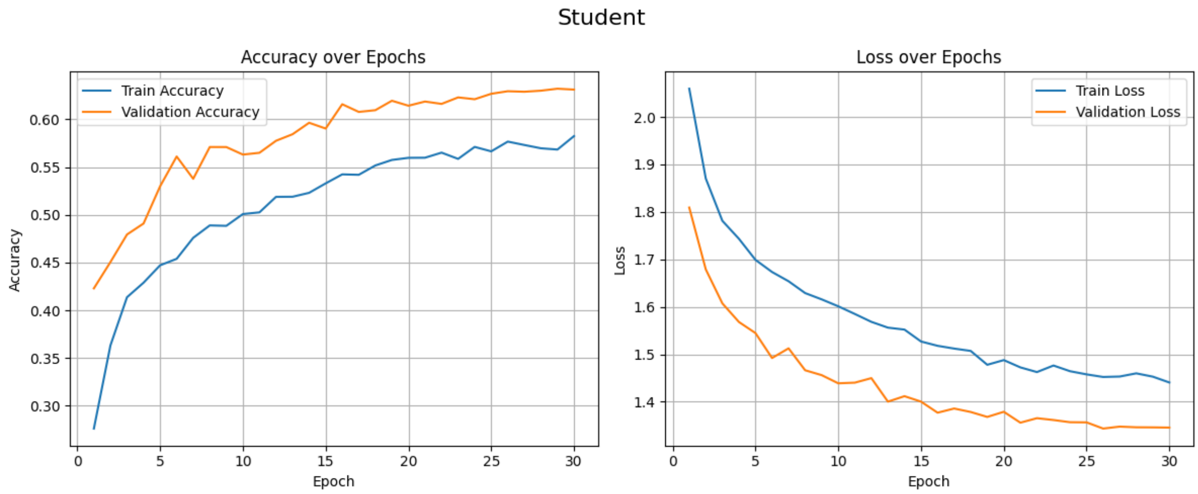
4.3 Task 4:

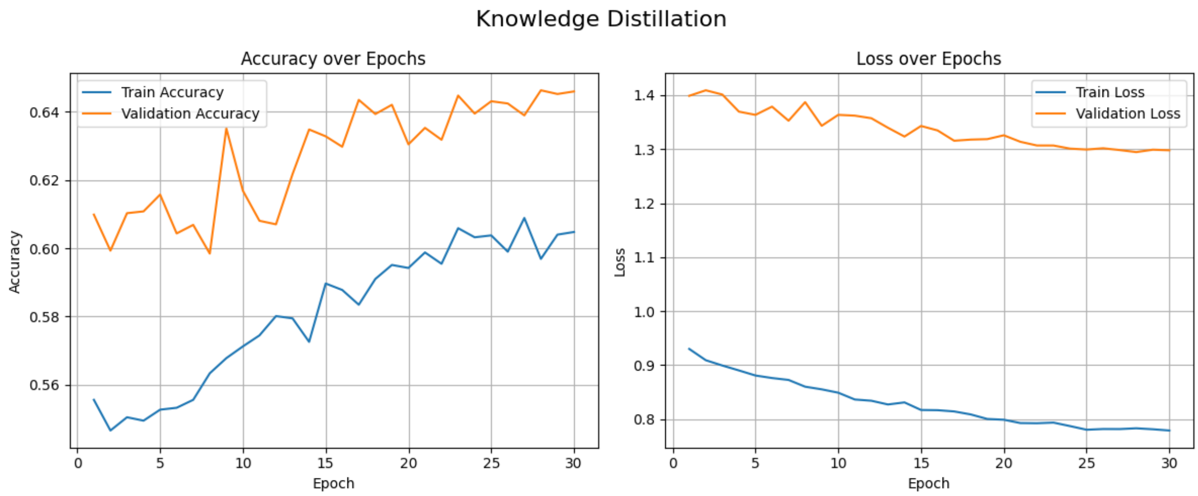
In here ResNet18 achieved the best performance with 81.1% validation accuracy and 0.95 loss. The small CNN trained from scratch reached only 62.8% accuracy with a higher loss of 1.35. Knowledge distillation improved the small CNN's accuracy to 65.7% and reduced the loss to 1.29. This confirms that distillation helps smaller models learn more effectively from larger ones.

Table 6: Performance Comparison for Task 4: Knowledge Transfer

Model	Validation Accuracy (%)	Validation Loss
ResNet18 (Teacher)	81.1	0.95
Small CNN (from scratch)	62.8	1.35
Small CNN (Distilled)	65.7	1.29







5 Discussion and Conclusion

This report compared CNN classifier performance on the CIFAR-10 dataset. A small CNN achieved an accuracy of 62.43%. Fine-tuning pre-trained models yielded superior results, with

ResNet50 reaching 80.2% and ResNet18 achieving 78.0% validation accuracy. Knowledge distillation significantly improved the small CNN's performance to 66.34% (distilled from ResNet50) and 65.7% (distilled from ResNet18), confirming its effectiveness in transferring knowledge to smaller, more efficient models.

Table 7: Model Performance and Notes

Model / Approach	**Accuracy (%)**	**Notes**
Small CNN (from scratch)	62.43	Baseline performance
ResNet50 (Fine-tuned)	80.2	Pre-trained, superior results
ResNet18 (Fine-tuned)	78.0	Pre-trained, superior results
Small CNN (Distilled - T3)	66.34	Improved with ResNet50 as teacher
Small CNN (Distilled - T4)	65.7	Improved with ResNet18 as teacher

6 Source Code

Project repository: https://github.com/Arif111866/Deep-Learning-AI